Probabilistic Grammars for Equation Discovery

Jure Brence[1,2], Nina Omejc[1,2], Boštjan Gec[1,2], Ljupčo Todorovski[3,1], and Sašo Džeroski[1,2]

(1) Jožef Stefan Institute, Department of Knowledge Technologies, Jamova 39, Ljubljana, Slovenia
(2) Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia
(3) University of Ljubljana, Faculty of Mathematics and Physics, Jadranska 21, Ljubljana, Slovenia

We focus on the problem of equation discovery, i.e., discovering closed-form equations from collections of numerical data. Generally, a model can be a single algebraic equation $y=f(x)$, an ordinary differential equation, or a system of ODEs. Our approach to equation discovery employs different types of probabilistic context-free grammars to encode knowledge on the mathematical modeling of systems in a domain of interest and generate candidate equations accordingly.

Our systems use probabilistic grammar as a generator of candidate expressions for the right-hand side of the equation and use numerical optimization to fit the values of the equation's constant parameters to the data. The system's output is a list of equations based on the expressions generated by the grammar, ranked by their degree of fit to the data. By specifying constraints on the space of candidate equations, a grammar provides a powerful formalism for expressing modeling knowledge. Furthermore, we can establish a map between the individual expression terms and the background knowledge concepts by analyzing the grammar parse tree of a particular expression.

We will illustrate how various aspects of background knowledge can be encoded with probabilistic grammars. The first is the parsimony principle, which prefers simpler equations over more complex ones. Many approaches would handle this principle with regularization, introducing parameters with infinite ranges: their optimal settings are typically sought in a series of computational experiments. In contrast, probabilistic context-free grammars parametrize parsimony through more comprehensible, limited-range parameters corresponding to rule probabilities. The generated expression's probability can also be used to infer the posterior distribution over the space of candidate equations or visualize it in a Pareto front of simple and accurate equations.

Furthermore, *probabilistic attribute grammars* allow us to encode the principles of dimensional analysis as a type of cross-domain knowledge that leverages knowledge of the measurement units of system variables to limit the space of candidate equations. We attach units as attributes to the grammar symbols, enabling the grammar rules to check the dimensional consistency of expressions. More complex attributes will enable the encoding of other concepts of modeling knowledge in the domain of interest.

We will present ProGED, an implementation of equation discovery with probabilistic grammars. Using different methods for parameter estimation, it supports the discovery of algebraic and ordinary differential equations, as well as exact integer equations from data. We will also present the results of applying ProGED to standard benchmarking problems in physics, as well as problems of modeling coupled oscillators of brain connectivity in neuroscience and integer sequences in mathematics.

Brence, J., Todorovski, L., and Džeroski, S. (2021) Probabilistic grammars for equation discovery. *Knowledge-Based Systems* 224: 107077. https://doi.org/10.1016/j.knosys.2021.107077

Gec, B., Omejc, N., Brence, J., Džeroski, S., and Todorovski, L. (2022) Discovery of Differential Equations Using Probabilistic Grammars. In *Proceedings of the 25th Conference on Discovery Science (pp. 22–31)*. Springer. https://doi.org/10.1007/978-3-031-18840-4_2

Nina Omejc, Boštjan Gec, Jure Brence, Ljupčo Todorovski, Sašo Džeroski (2023). Probabilistic grammars for modeling dynamical systems from coarse, noisy, and partial data. *In review.* https://doi.org/10.21203/rs.3.rs-2678362/v1

Brence, J., Džeroski, S., and Todorovski, L. (2023) Dimensionally consistent equation discovery through probabilistic attribute grammars. *Information Sciences.* https://doi.org/10.1016/j.ins.2023.03.073